TRANSCRIPTION OF BROADCAST TELEVISION AND RADIO NEWS: THE 1996 ABBOT SYSTEM

G.D. Cook *D.J. Kershaw* *J.D.M. Christie* *A.J. Robinson*

Cambridge University Engineering Department,
Trumpington Street,
Cambridge, CB2 1PZ, UK

ABSTRACT

ABBOT is a hybrid connectionist-HMM large vocabulary continuous speech recognition system developed at the Cambridge University Engineering Department. This uses a recurrent neural network acoustic model to map acoustic features into posterior phone probabilities. These posterior probabilities are then converted to scaled likelihoods and used as observation likelihoods for phone HMMs [1, 2]. This paper describes the development of the CU-CON system which participated in the 1996 ARPA Hub 4 Evaluations. The system is based on ABBOT. The Hub 4 Evaluation task involves the transcription of broadcast television and radio news programmes. This is an extremely demanding task for state-of-the-art speech recognition systems. Typical programmes include a wide variety of speaking styles and acoustic conditions. These range from read speech recorded in the studio to extemporaneous speech recorded over telephone channels. Results are presented for the system at various stages of development, as well as for the final evaluation system.

1. INTRODUCTION

The hybrid connectionist-hidden Markov model approach uses an underlying hidden Markov process to model the time-varying nature of the speech signal and a connectionist system to estimate the observation likelihoods within the hidden Markov model (HMM) framework. ABBOT is a large-vocabulary speech recognition system based on the hybrid approach which utilises a recurrent network for acoustic modelling. The major advantage of this approach is that the recurrent network acts as a non-parametric model that is able to capture temporal acoustic context. Consequently, the basic ABBOT system is able to achieve very good performance using single pass decoding and context-independent phone models [3].

This paper reports on the development of the CU-CON system for the 1996 ARPA Evaluations. Section 3 describes the acoustic models used for the 1996 evaluations, and the process of training a new set of models on the broadcast news acoustic training data. This includes a description of the linear input network (LIN) technique used for channel adaptation. This method has been used to adapt the acoustic models used for telephone speech, and for speech in degraded acoustical conditions. Section 4 outlines the procedure used for creating a lexicon and language model, plus a description of the training texts used, and the procedure for producing pronunciations. Next the performance of the system at various stages of development is assessed on the 1996 Hub 4 development test data. The final section presents the official results on the Hub 4 evaluation test data.

2. THE 1996 ARPA HUB 4 TASK

The 1996 evaluation consists of two components, a "partitioned evaluation" (PE) component, and an "unpartitioned evaluation" (UE) component. The PE contains speech that is manually segmented into homogeneous regions, and provides a set of six controlled contrastive conditions known as "evaluation focus conditions":

F0: Baseline broadcast studio speech

F1: Spontaneous broadcast studio speech

F2: Speech over telephone channels

F3: Speech in the presence of background music

F4: Speech under degraded acoustical conditions

F5: Speech from non-native speakers

Segments that do not fall within the specification for the focus conditions presented above are labelled FX. The UE is similar to the 1995 Hub 4 evaluation in that it contains relatively complete portions of television and radio news broadcasts, but using a wider variety of source material than was employed in the 1995 evaluation. The CU-CON system participated in the PE only.

3. ACOUSTIC MODELS

This section describes the acoustic modelling process used in the ABBOT system. This includes a brief description of the front-end, the recurrent network, and phonetic context-dependent modelling which augments the standard context-independent model.

3.1. Acoustic Feature Representation

Two sets of acoustic features have been used in the past by the ABBOT system: MEL+, a 20 channel mel-scaled filter bank with energy, degree of voicing, and pitch [4], and PLP, 12th order cepstral coefficients derived using perceptual linear prediction and log energy [5]. The 1996 ABBOT system uses both MEL+ and PLP acoustic features. The MEL+ and PLP features were computed from 32 msec windows of the speech waveform every 16 msec. To increase the robustness of the system to environmental conditions, the statistics of each feature channel were normalised to zero mean with unit variance over each segment.

3.2. Acoustic Model Architecture

The basic acoustic modelling system [6, 7] is illustrated in Figure 1. For each input frame, an acoustic vector, $\mathbf{u}(t)$, is presented at the input to the network along with the current state, $\mathbf{x}(t)$. These two vectors are passed through a standard single layer, feed-forward network to give the output vector, $\mathbf{y}(t-4)$, and the next state vector, $\mathbf{x}(t+1)$. Sigmoid and softmax nonlinearities are applied to the state and output nodes, respectively. The output vector represents an estimate of the posterior probability of each of the phone classes, i.e.,

$$y_i(t) \simeq \Pr(q_i(t)|\mathbf{u}_1^{t+4}) \tag{1}$$

where $q_i(t)$ is state i at time t and $\mathbf{u}_1^t = \{\mathbf{u}(1), \ldots, \mathbf{u}(t)\}$ is the input from time 1 to t. The output is delayed by four frames to account for forward acoustic context. The state vector provides the mechanism for modelling acoustic context and the dynamics of the acoustic signal. There is one output node per phone and the recurrent network generates all the frame-by-frame phone posterior probabilities in parallel.

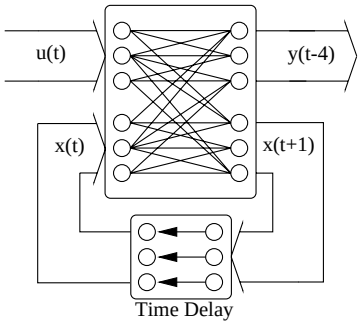

Figure 1: The recurrent network used for phone probability estimation.

The training approach is based on Viterbi training. Each frame of training data is assigned a phone label based on an utterance orthography and the current model. The recurrent network is then trained – using the back-propagation-through-time algorithm [8] – to map the input acoustic vector sequence to the phone label sequence. The labels are then reassigned and the process iterates. Initial alignments for the ABBOT system were derived from a recurrent network trained on the TIMIT database.

The 1996 ABBOT system utilises recurrent networks trained on forward-in-time and backward-in-time input sequences of both the MEL+ and PLP feature vectors. The recurrent network builds up a representation of the past acoustic context which implies the ordering of the input data is important. A significant performance improvement is achieved by merging multiple recurrent networks trained on these different input representations [9]. The most successful merging technique merges the network outputs in the log domain, i.e.,

$$\log y_i(t) = \frac{1}{K} \sum_{k=1}^{K} \log y_i^{(k)}(t) - Z \tag{2}$$

where Z is a constant to insure that $\mathbf{y}$ is a valid probability distribution.

3.3. Context-Dependent Modelling

By using the definition of conditional probability, the factorisation of conditional context-class probabilities is used to implement phonetic context-dependency in the acoustic model [10]. The joint posterior probability of context class j and phone class i is given by,

$$y_{ij}(t) = y_i(t) y_{j|i}(t), \tag{3}$$

where $y_i(t)$ is estimated by the recurrent network. Single-layer networks or "modules" are used to estimate the conditional context-class posterior,

$$y_{j|i}(t) \simeq \Pr(c_j(t)|\mathbf{u}_1^{t+4}, q_i(t)), \tag{4}$$

where $c_j(t)$ is the context class for phone class $q_i(t)$. The input to each module is the internal state (similar to the hidden layer of an MLP) of the recurrent network, since it is assumed that the state vector contains all the relevant contextual information necessary to discriminate between different context classes of the same monophone [11, 12].

Figure 2 shows the context-dependent system in operation. The outputs on the right hand side of this figure are the context-dependent posterior probabilities as estimated by Equation 3.

Viterbi segmentation is used to align the training data. Each context network is trained on a non-overlapping subset of the state vectors generated from all the Viterbi aligned training data. The context networks are trained using a gradient-based procedure. The context classes for each context module are

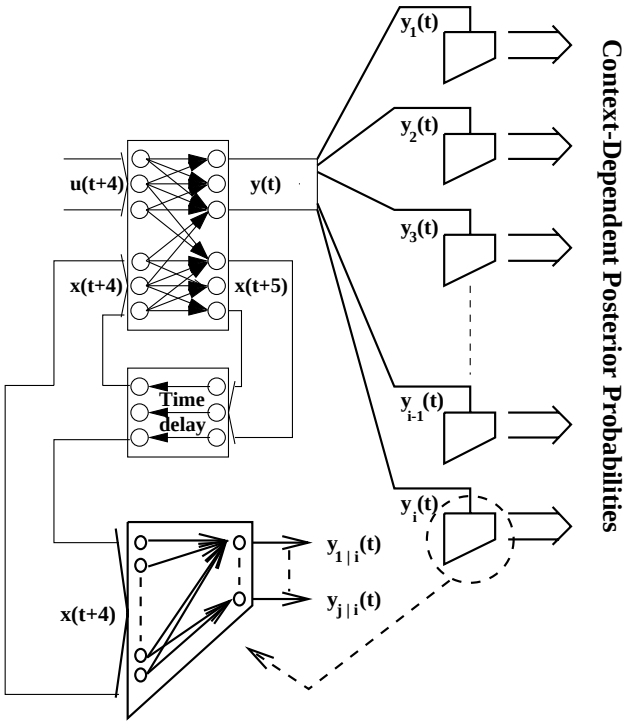

Figure 2: The phonetic context-dependent recurrent neural network modular system.

determined by using a decision tree based approach. This allows for sufficient statistics for training and keeps the system compact (allowing fast context training). The decision trees are also used to relabel the pronunciation lexicon.

3.4. Acoustic Model Training

This section describes the development of the acoustic models used in the 1996 ABBOT system.

A Viterbi forced alignment was performed using the 1995 ABBOT acoustic models. These are forward and backward in time PLP models trained on the secondary channel data from the Wall Street Journal corpus (SI84). Average log probability scores were generated for each segment. Those segments with poor scores were checked manually. It was found necessary to edit the transcriptions or time markings for approximately 2.5% of the segments.

A forward and backward PLP model was then trained on all of the broadcast news data. Forward and backward MEL+ models were also trained on this data. Only one Viterbi alignment was performed due to the late arrival of the acoustic training data, and the lack of time available. These models are denoted `BN`. A further 4 acoustic models were trained solely on the F0 segments. These comprise forward and backward in time models for both MEL+ and PLP, and are denoted `BN.F0`.

3.5. Channel Adaptation

The `BN` models were extended to the F2 and F4 conditions by means of linear input network (LIN) adaptation on the training data. The linear input network (LIN) has been successfully applied to connectionist HMM hybrid systems for supervised speaker adaptation [13], unsupervised speaker adaptation [14], and unsupervised channel adaptation [3, 15]. A linear mapping is created to transform the acoustic vector. During recognition, this transformed vector is fed as input to the speaker independent RNN. To train the LIN for a new focus condition, the LIN's weights are initialised to an identity matrix; this guarantees that the initial starting point is the general broadcast news model. The input is propagated forward to the output layer of the RNN. At this point the error is back-propagated through the RNN. Note that the RNN weights are kept frozen, and only the LIN's weights are updated.

The F2 data was marked as either having low or medium fidelity. We reclassified all the F2 data into narrow or wide band data based on the power in the upper 4kHz of the spectrum. However, merely averaging the power in the upper 4kHz of a segment would bias the classification due to the relative number of voiced and unvoiced sections in a segment. To account for this we multiplied the energy in the upper 4kHz of each frame by the estimated probability of the frame representing an unvoiced segment. We chose a threshold for the choice of narrow bandwidth and full bandwidth by manually classifying a small proportion of the F2 segments. After setting this threshold all the F2 segments were relabelled. A LIN was trained for each `BN` model on the narrow bandwidth F2 data. These adapted models (denoted `BN.adpt-nb`) were used on the evaluation data classified as narrow bandwidth. Those segments classified as F2 wideband were recognised using the `BN` model set without adaptation.

For the F4 condition LIN networks were trained on those segments labelled as F4 in the training data. These models are denoted `BN.adpt-F4`.

4. LANGUAGE MODEL AND LEXICON

The 1996 ABBOT system uses a 65,532 word vocabulary. This was produced by extracting the most frequent 80,000 words from the broadcast news text data only, and removing misspelled words, processing errors etc. Trigram language models were built using an alpha release of the CMU-Cambridge Statistical Language Modelling Toolkit version 2.0. The toolkit offers more efficient processing of text data, and provides for various discounting strategies [16]. More details of the CMU-Cambridge Statistical Language Modelling Toolkit version 2.0 can be found at `http://svr-www.eng.cam.ac.uk/~prc14/toolkit_documentation.html`. The language models used by ABBOT for previous evaluations have used the Good-Turing discounting method. However, this year's language

models have used the Witten-Bell discounting method [17].

Initial experiments were performed using both the broadcast news texts, and the 1995 Hub 4 data, which covers general North American business news. The results of these experiments can be seen in Table 1.

Focus	OOV	Perplexity	
		BN Texts	BN + Hub 4
F0	0.76%	210.06	193.62
F1	0.50%	194.70	206.68
F2	0.53%	190.50	196.52
F3	1.14%	238.03	230.87
F4	0.71%	225.85	214.34
F5	0.98%	299.15	252.16
FX	0.57%	197.97	206.45
All	0.65%	206.05	203.07

Table 1: Perplexity and out-of-vocabulary (OOV) rate by focus on the acoustic training data transcriptions for two language models, one trained on the broadcast news texts, and one trained on both the broadcast news texts and the 1995 Hub 4 texts.

From the initial results it was decided to build two different language models, one for speech considered "planned", and one for speech considered "spontaneous". Table 2 shows the different text sources for the language model training data. The Marketplace data is the transcriptions of the training data supplied for the 1995 Hub 4 Evaluation. The transcriptions of the broadcast news acoustic training data were also used for training the language models.

Texts	N°. Words	Language Model
Broadcast News	132 million	planned, spont.
1995 Hub 4 texts	108 million	planned.
1995 Marketplace	50,000	planned, spont.
1996 transcripts	380,000	planned, spont.

Table 2: LM training data.

The recognition lexicon includes priors on multiple pronunciations. The priors are normally calculated by gathering the statistics from a forced alignment. This year these multiple pronunciation priors have been reestimated (and smoothed with the statistics from the standard forced alignment), for spontaneous speech. The statistics were gathered from a forced alignment on a phone string recognition of the F1 and F2 training data.

5. RESULTS

Table 3 shows results on the development test data for various systems. These systems represent various stages in the development of the 1996 ABBOT system:

Focus	Word Error Rate %		
	System 1	System 2	System 3
F0	31.9	22.9	18.8
F1	58.0	46.8	40.9
F2	66.6	51.6	45.7
F3	62.9	46.6	40.7
F4	48.2	33.8	27.4
F5	44.7	36.6	31.5
FX	73.0	61.7	58.1
Overall	54.6	42.7	37.5

Table 3: Results on the development test set.

System 1: This is based on the 1995 ABBOT system, except that only a single forward context-independent PLP network was used. The acoustic model training data is the short term speakers from WSJ0 secondary channel (SI84). The standard ARPA 1995 60,000 word trigram language model was used.

System 2: This system uses forward and backward PLP broadcast news context-independent acoustic models. A trigram language model trained only on the broadcast news text is used. The system has a 65,532 word vocabulary.

System 3: This system uses word-internal context-dependent forward and backward PLP acoustic models. The same language model as **system 2** was used.

From the results it can be seen that using the broadcast news acoustic and language modelling training data, and merging forward and backward acoustic models has resulted in a significant reduction in error rates. The overall error rate has been reduced from 54.6% to 42.7%, a reduction of 22.8%. The addition of limited word internal context-dependent models has further reduced the overall word error rate to 37.5%, a improvement in performance of 12.2%. Note that the adapted models `BN.adpt-nb` or `BN.adpt-F4` have not been evaluated on the development data due to lack of time. The models trained only on those segments marked F0 (`BN.F0`) result in a word error rate of 16.2% on the F0 segments of the development test set, a reduction of 13.8% compared to the `BN` models.

6. EVALUATION SYSTEM

The CU-CON evaluation system used a number of features that were not used on any of the systems evaluated on the development data. Different language models were used for segments marked as planned speech and segments marked as

spontaneous speech. In addition, channel adaptation was used for reduced bandwidth F2 segments, and for the F4 segments. Side information indicating planned or spontaneous speech is provided with the FX segments. This information was used to select the appropriate acoustic and language model to use for each of the FX segments. Table 4 lists the acoustic and language models used for each of the segments in the evaluation test data. Note that the narrow band, wide band classification of the F2 and FX-F2 segments was accomplished using the method described in Section 3.5.

Focus	Acoustic Model	Language Model
F0	BN.F0	planned
F1	BN	spont.
F2.nb	BN.adpt-nb	spont.
F2.wb	BN	spont.
F3	BN	planned
F4	BN.adpt-F4	planned
F5	BN	planned
FX-F1	BN	spont.
FX-F2.nb	BN.adpt-nb	spont.
FX-F2.wb	BN	spont.

Table 4: Acoustic and language models used for the various focus conditions.

Table 5 shows the official word error rates of the CU-CON system on the 1996 Hub 4 evaluation test data. The number of words per focus condition is also included. Single pass decoding was performed using the NOWAY decoder [18]. No test set adaptation was performed for this evaluation.

Focus	N$^{o.}$ Words	WER %
F0	5995	25.8
F1	6593	33.5
F2	1748	40.4
F3	1417	33.4
F4	1833	39.3
F5	299	40.5
FX	2301	53.1
Overall	20186	34.7

Table 5: Number of words and word error rate by focus for the CU-CON evaluation system.

Comparison with the results in Table 3 shows that error rates for the baseline F0 condition are significantly higher on the evaluation data. The perplexity of the F0 segments of the development and evaluation data is similar, as is the signal-to-noise ratio (SNR) (27.6dB for the evaluation data, and 29.4dB for the development data). It is therefore surmised that the F0 evaluation data contains more conversational type speech than it's development counterpart.

Focus	Perplexity	OOV Rate %
F0	205.28	1.59
F1	120.57	1.40
F2	150.22	1.63
F3	285.18	1.53
F4	128.20	0.59
F5	271.14	0.33
FX	167.29	0.96

Table 6: Perplexity and OOV by focus for the CU-CON evaluation system.

The perplexity for the different focus conditions is shown in Table 6. The F0 perplexity is considerably higher than seen for read speech in previous evaluations. Typical perplexity values for the 1995 Hub 3 Evaluation test data were in the region of 130 for trigram language models. Another possible reason for the high error rates for planned speech when compared with previous read speech evaluations, may be the low signal-to-noise ratio (SNR). The SNR of the F0 segments is 27.6dB (as measured by the NIST tool `wavemd`), compared to 38.0dB for the clean read speech of the 1995 Hub 3 Evaluation contrast.

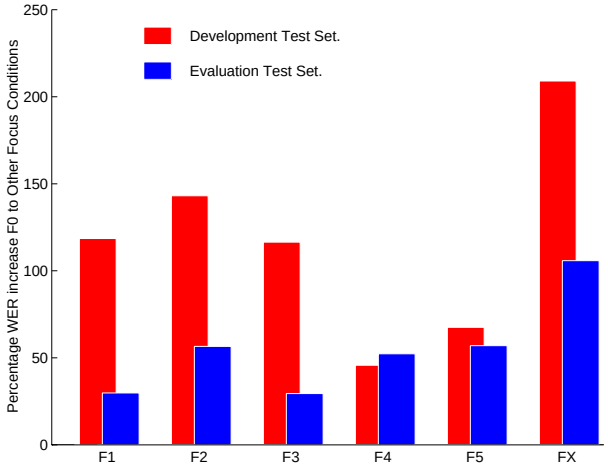

Figure 3: Relative WER increase from the baseline F0 focus condition to each of the other focus conditions.

Figure 3 shows the degradation in performance of the different focus conditions as measured against the baseline F0 focus, for both the development and evaluation test data. It can be seen that a far greater degradation was observed on the development data, however, this is likely to reflect the significantly lower word error rate of F0. The relative degradation between each of the focus conditions is similar for the

development and evaluation data, except for the F4 and F5 focus conditions. These exhibit a far greater degradation on the evaluation data when compared with the other focuses. Investigation has revealed that the SNR of the F4 data is 25.1dB for the development data, but only 18.6dB for the evaluation data. This is likely to be the reason for the greater F4 degradation. The source of the extra degradation seen for the F5 focus condition is most probably due to the far higher perplexity seen in the evaluation data, which is 28% higher than on the development data.

7. CONCLUSIONS

This paper has described the development of the CU-CON system for the recognition of broadcast television and radio news. This has concentrated on building acoustic and language models on data from this domain. This approach was necessitated by the late arrival of the training data. Further work on this task is planned, and includes the use of boosting [19], extended context-dependent modelling, test set adaptation, and speech enhancement.

8. ACKNOWLEDGEMENTS

This work was partially funded by ESPRIT project 20007 SPRACH. Thanks to Steve Renals for his help in producing evaluation results. The authors also acknowledge the help of Rachel Morton and Rishi Nag in tidying of training data transcriptions and in the production of pronunciations.

References

1. H.A. Bourlard and N. Morgan. *Connectionist Speech Recognition: A Hybrid Approach.* Kluwer Acedemic Publishers, 1994.
2. M.M. Hochberg, G.D. Cook, S.J. Renals, A.J. Robinson, and R.S. Schechtman. The 1994 ABBOT Hybrid Connectionist-HMM Large Vocabulary Recognition System. *Proceedings of the ARPA Spoken Language Systems Technology Workshop*, 1995.
3. D.J. Kershaw, S. Renals, and A.J. Robinson. The 1995 ABBOT LVCSR System for Multiple Unknown Microphones. In *Int. Conf. in Spoken Language Processing*, October 1996.
4. A.J. Robinson. Several Improvements to a Recurrent Error Propagation Network Phone Recognition System. Technical Report CUED/F-INFENG/TR.82, Cambridge University Engineering Department, September 1991.
5. H. Hermansky and N. Morgan. RASTA Processing of Speech. *IEEE Transactions on Speech and Audio Processing*, 2(4):578–89, October 1994.
6. A.J. Robinson. An Application of Recurrent Nets to Phone Probability Estimation. *IEEE Transactions on Neural Networks*, 5(2):298–305, March 1994.
7. A.J. Robinson, M.M. Hochberg, and S.J. Renals. The Use of Recurrent Neural Networks in Continuous Speech Recognition. In C. H. Lee, K. K. Paliwal, and F. K. Soong, editors, *Automatic Speech and Speaker Recognition – Advanced Topics*, chapter 19. Kluwer Academic Publishers, 1995.
8. P.J. Werbos. Backpropagation Through Time: What Does It Mean and How to Do It. In *IEEE*, volume 78, pages 1550–60, October 1990.
9. M.M. Hochberg, G.D. Cook, S.J. Renals, and A.J. Robinson. Connectionist Model Combination for Large Vocabulary Speech Recognition. In *Neural Networks for Signal Processing*, volume IV, pages 269–278, 1994.
10. H. Bourlard and N. Morgan. Continuous Speech Recognition by Connectionist Statistical Methods. *IEEE Transactions on Neural Networks*, 4(6):893–909, November 1993.
11. D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. In D.S. Touretzky, M.C. Mozer, and M.E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, Cambridge, MA 02142-1399, 1996.
12. D.J. Kershaw, M.M. Hochberg, and A.J. Robinson. Incorporating Context-Dependent Classes in a Hybrid Recurrent Network-HMM Speech Recognition System. F-INFENG TR217, Cambridge University Engineering Department, May 1995.
13. J. Neto, L. Almeida, M.M. Hochberg, C. Martins, L. Nunes, S.J. Renals, and A.J. Robinson. Speaker Adaptation for Hybrid HMM-ANN Continuous Speech Recognition Systems. In *Eurospeech*, pages 2171–2174, September 1995.
14. J.P. Neto, C.A. Martins, and L.B. Almeida. Unsupervised Speaker-Adaptation For Hybrid HMM-MLP Continuous Speech Recognition System. In *IEEE Speech Recognition Workshop*, pages 187–8, December 1995.
15. D.J. Kershaw, A.J. Robinson, and S.J. Renals. The 1995 Hybrid Connectionist-HMM Large-Vocabulary Recognition System. In *ARPA Speech Recognition Workshop*, Harriman House, New York, February 1996.
16. P.R Clarkson and R. Rosenfeld. Statistical Language Modelling with the CMU-Cambridge Toolkit. Submitted to EuroSpeech 1997.
17. I.H. Witten and T.C. Bell. The Zero-Frequency Problem: Estimating the Probabilities of Novel Events in Adaptive Text Compression. *IEEE Transactions on Information Theory*, 37(4):1085–1094, 1991.
18. S.J. Renals and M.M. Hochberg. Decoder Technology for Connectionist Large Vocabulary Speech Recognition. Technical Report CS-95-17, Dept. of Computer Science, University of Sheffield, 1995.
19. G.D. Cook and A.J. Robinson. Boosting the Performance of Connectionist Large Vocabulary Speech Recognition. In *Int. Conf. in Spoken Language Processing*, 1996.